NIST LoReHLT 2016 Evaluation Plan

Last Updated May 20, 2016

1 Introduction

The DARPA Low Resource Languages for Emergent Incidents (LORELEI) Program seeks to develop human language technology (HLT) that can support rapid and effective response to emerging incidents where the language resources are very limited. As such, LORELEI aims to develop capabilities that can extract knowledge from foreign language sources quickly. This document describes the evaluation specifications of the component evaluation conducted by NIST to assess the performance and track the progress made.

Participation in the NIST Low Resource Human Language Technology (LoReHLT) evaluation is required for all DARPA LORELEI performers responsible for the relevant component technologies in LORELEI. The evaluation is also open to all researchers who find the evaluation tasks of interest. There is no cost to participate. However, participants are expected to attend a post-evaluation workshop to present and discuss their systems and results at their own expense. Information and updates about the component evaluation will be posted to the NIST LoReHLT website[1].

2 Evaluation Tasks

There are three evaluation tasks. LORELEI performers are required to participate in the tasks as outlined by their Statement of Work. Open participants can participate in any and all tasks.

- **Machine Translation (MT)** – for each document, automatically translate it from a given incident language (IL) to English. For MT specific requirements, see Section 0.
- **Situation Frame (SF)** – for each document, automatically generate Situation Frames covered in the document. For SF specific requirements, see Section 13
- **Named Entity Recognition (NER)**[2] – for each document, identify and classify named mentions of PER, GPE, ORG, LOC entities. For NER specific requirements, see Section 14.

3 Training Conditions

For each evaluation task, there are two training conditions (constrained and unconstrained) that differentiate the amount/source of incident language-related training material without preventing/excluding multilingual resources and technology. The intent of the *constrained* training condition is to test multilingual systems that are re-targeted to an incident language using a fixed amount of incident language materials. Teams should consult with NIST if their approach is not easily classifiable.

[1] http://www.nist.gov/itl/iad/mig/lorehlt16.cfm

[2] This task is for year 1 only. In subsequent years (2+), the task will be Entity Discovery and Linking (EDL).

- **Constrained** – The constrained data condition limits the incident language material used to train/adapt the tested technology to only those distributed according to Section 5 (IL Data) and Section 6 (Native Language Informants). No other incident language materials, i.e., parallel text, speech corpora, etc. are permitted but knowledge gained from the Native Language Informant is permitted. Prior to the evaluation period, which begins with the announcement of the IL, teams can assemble multilingual resources/technologies/etc. to use during the evaluation so long as they are multilingual-focused in nature. Serendipitous included incident language data in a multilingual system is allowed and must be documented in the system description. The use of mono- and bi-lingual resources is allowed so long as they do not include the incident language. The Constrained training condition is **required for each task participated**.
- **Unconstrained** – The unconstrained condition removes the limitations of the constrained condition. Teams can use additional, publicly available, incident language materials obtained before or after the IL announcement from an epoch before or after the incident. Teams can use pre-existing, mono-lingual technologies for the incident language. Teams can use additional Native Language Informant time beyond the limits in Section 6. The teams must document the additional data and technologies in the system description. The unconstrained training condition **optional but encouraged**.

4 Baseline Training Data

For each evaluation task, a set of non-IL data resources will be provided by the LDC for training prior to the evaluation period. To obtain this data, open participants must register to participate and sign the license agreement which can be found on the NIST LoReHLT website.

Each task (MT, SF, or NER) has its own annotation guideline. If you are an open participant and do not have direct access to the annotation guidelines, please contact LDC at lorelei-poc@ldc.upenn.edu and ask for the LoReHLT translation, situation frame, or simple named entity guidelines.

5 Evaluation Data

5.1 Component Definition & Release Plan

All three evaluation tasks will use the same data component and have the same release plan. The LDC releases the Incident Language (IL) data and English Scenario Model in an encrypted format (see 5.4), and NIST releases the appropriate decryption key(s) at the appropriate stages. Participants must complete an ensemble of all three checkpoints for their submissions to be considered complete. The stages are:

- Pre-IL Announcement (before the IL Announcement)
  - **Set 0**: Encrypted pre-incident IL training data released
  - **Set 1**: Encrypted incident/post-incident IL training data set 1 released
  - **Set 2**: Encrypted incident/post-incident IL training data set 2 released
  - **Set S**: Encrypted incident/post-incident English Scenario Model released
  - **Set E**: Encrypted incident/post-incident IL evaluation data released
- IL Announcement

  - Identity of IL announced
  - Decryption keys for **set 0** and **set E** released
- Evaluation Checkpoint 1
  - Train with data from **set 0** begins at IL Announcement
  - Evaluation Checkpoint 1 submission due 7 days after IL Announcement
  - Decryption key for **set 1** released 7 days after IL Announcement and after submission to Evaluation Checkpoint 1 made
- Evaluation Checkpoint 2
  - Train with data from **set 0** begins at IL Announcement
  - Train with data from **set 1** begins after the Evaluation Checkpoint 1 submission deadline and the team makes a submission
  - Evaluation Checkpoint 2 submission due 14 days after IL Announcement
  - Decryption key for **set 2** released 14 days after IL Announcement and after submission to Evaluation Checkpoint 2 made
- Evaluation Checkpoint 3
  - Train with data from **set 0** begins at IL Announcement
  - Train with data from **set 1** begins after the Evaluation Checkpoint 1 submission deadline and the team makes a submission
  - Train with data from **set 2** and **set S** begins after the Evaluation Checkpoint 2 submission deadline and the team makes a submission
  - Evaluation Checkpoint 3 submission due 30 days after IL Announcement

5.2 Data Description

The composition of the five datasets (**set 0, set 1, set 2, set S, and set E**) are listed in Table 1 below. The given target data volume is **approximate** and depends on data availability. If the amount for a genre is short of the target, LDC will substitute with another genre. “Kw” refers to multiples of 1000 words.

5.3 Data Format and Structure

These five datasets (aka the evaluation IL package) will be released by the LDC. The data format and structure are described in detail in the data specification document uploaded on the NIST LoReHLT website.

5.4 Data Encryption

The dataset described above will be encrypted using OpenSSL. NIST has created a package with instructions on how to encrypt and decrypt the data using some sample data. The package can be downloaded from the NIST LoReHLT website.

Table 1: LoReHLT16 IL data description

Set 0 – pre-incident epoch
Category I Resources[3] • Monolingual Source Text: ○ ~100Kw newswire ○ ~75Kw discussion forum/blog ○ ~50Kw Twitter/SMS • Parallel Text[4]: ○ ~100Kw newswire ○ ~100Kw discussion forum/blog ○ ~100Kw Twitter/SMS • Parallel Dictionary (~10,000 stems/lemmas) Category II Resources (any 5 of the following): • parallel dictionary IL --> non-English • monolingual IL dictionary • monolingual IL grammar book • parallel English --> IL grammar book • monolingual IL primer book • monolingual IL gazetteer • parallel IL --> English gazetteer
Set 1 – incident/post-incident epoch
Monolingual Source Text – 1/3 of leftover after **set E** is met
Set 2 – incident/post-incident epoch
Monolingual Source Text – 2/3 of leftover after **set E** is met
Set S – incident/post-incident epoch
English Scenario Model – approximately 50Kw, genre balance will vary based on availability
Set E – incident/post-incident epoch
Source Text: • ~100Kw newswire • ~50Kw discussion forum/blog • ~50Kw Twitter/SMS

6 Native Informant Resources

During the evaluation period, participants are allowed the use of a native informant (NI) in their system development. The LORELEI performers will be provided the native informant by their sponsor[5] through

[3] One of the category I resources (monolingual text, parallel text, or parallel dictionary) must exceed the minimum target by 500%.

[4] The parallel text is found/harvested data and automatically aligned, not created (e.g. via professional translation agency or crowdsourcing). ~300Kw comparable may be substituted for every 100Kw parallel if parallel text is not available.

the data provider Appen. The native informant will be available remotely via telephone or internet connection. Open participants, if they wish to use a native informant, have to supply their own at their own cost and are free to determine how they communicate with their informant. However, consultation with the informant, by LORELEI performers and open participants, must abide by the following guidelines:

- Informant can be a native speaker of the IL but cannot be a professional linguist.
- It is up to the individual teams to determine how they will make use of the informant. However, **the evaluation data must remain unseen and sequestered, and all probings of the evaluation data are prohibited**. The teams must document how they have used the informant (e.g. producing additional resources for training, etc.).
- If a member(s) of the developer's team also happens to be a native speaker of the IL, this information must also be documented.
- For the constrained training condition, consultation with the informant is limited to the number of hours listed below for each task a team participates regardless of how many submissions. If the use of the native informant exceeds the number of hours given, the submissions are considered to be in the unconstrained training track.
    - 1 hour for Evaluation Checkpoint 1
    - 5 hours for Evaluation Checkpoint 2 (4 hours if 1 hour was used in Checkpoint 1)

7 Evaluation Protocol

7.1 Evaluation Account

All participants are required to sign up for an evaluation account on the NIST LoReHLT evaluation web site as all evaluation activities will be conducted via the evaluation account. Go to https://lorehlt.nist.gov to sign up for an account. Participants will need a valid email address and choose a password that is at least 12 characters long including uppercase and lowercase letters, numbers, and special characters.

After signing up and confirming the account, each participant[6] will be asked to associate himself/herself to a site[7] (or create his/her site if it does not exist). The first person who creates the site will be deemed the *site representative* and will have to approve participants who want to join his/her site. The site representative will be asked to associate his/her site to a team[8] (or create his/her team if it does not exist). The first person who creates the team will be deemed the *team representative* and will have to approve sites who want to join his/her team. The site representative can create multiple teams as well as ask to join his/her site to other teams. The team representative must register his/her team for a particular task to participate in that task. If the site declares itself as a LORELEI performer, its status will

[5] LORELEI performers will be provided NI time by their sponsor only for the amount given above. If teams want additional time, they must make their own arrangement at their own cost.
[6] A *participant* is defined as a member of an organization who takes part in the evaluation (e.g., Clark Kent).
[7] A *site* is defined to be a single organization participating in the evaluation (e.g., The Daily Planet).
[8] A *team* is defined to be a group of organizations collaborating on a task in the evaluation (e.g., Justice League).

be verified. If the site is not a LORELEI performer, the site representative will be asked to sign the LDC license. The LDC will confirm the license and release the appropriate data to the site.

7.2 System Input File Format

The input source data to the system is the same across all three tasks and uses the LDC LTF format conforming to the LTF DTD referenced inside the test files. For a detailed description of the evaluation IL package, see Section 5.3.

Each team is to process the entire test set even though for some tasks only a portion of the test will be scored. An example LTF file is given below.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE LCTL_TEXT SYSTEM "ltf.v1.5.dtd">
<LCTL_TEXT>
  <DOC id="NW_ARX_UZB_164780_20140900" tokenization="tokenization_parameters.v2.0" grammar="none"
raw_text_char_length="1781" raw_text_md5="1511bf44675b0256adc190a7b96e14bd">
    <TEXT>
      <SEG id="segment-0" start_char="0" end_char="31">
        <ORIGINAL_TEXT>Emlashni birinchi kim boshlagan?</ORIGINAL_TEXT>
        <TOKEN id="token-0-0" pos="word" morph="none" start_char="0" end_char="7">Emlashni</TOKEN>
        <TOKEN id="token-0-1" pos="word" morph="none" start_char="9" end_char="16">birinchi</TOKEN>
        <TOKEN id="token-0-2" pos="word" morph="none" start_char="18" end_char="20">kim</TOKEN>
        <TOKEN id="token-0-3" pos="word" morph="none" start_char="22" end_char="30">boshlagan</TOKEN>
        <TOKEN id="token-0-4" pos="punct" morph="none" start_char="31" end_char="31">?</TOKEN>
      </SEG>
      <SEG id="segment-1" start_char="33" end_char="61">
        <ORIGINAL_TEXT>Pereyti k: navigatsiya, poisk</ORIGINAL_TEXT>
        <TOKEN id="token-1-0" pos="word" morph="none" start_char="33" end_char="39">Pereyti</TOKEN>
        <TOKEN id="token-1-1" pos="word" morph="none" start_char="41" end_char="41">k</TOKEN>
        <TOKEN id="token-1-2" pos="punct" morph="none" start_char="42" end_char="42">:</TOKEN>
        <TOKEN id="token-1-3" pos="word" morph="none" start_char="44" end_char="54">navigatsiya</TOKEN>
        <TOKEN id="token-1-4" pos="punct" morph="none" start_char="55" end_char="55">,</TOKEN>
        <TOKEN id="token-1-5" pos="word" morph="none" start_char="57" end_char="61">poisk</TOKEN>
       </SEG>
       ...
    </TEXT>
  </DOC>
</LCTL_TEXT>
```

7.3 System Output File Format

While all tasks have the same system input file format, each has its own output format. Refer to the task specific section for information about the output requirement for that task.

7.4 File List

The terms of usage of the Twitter data require that only the URLs of the tweets can be redistributed, not the actual tweets. Tweets can be deleted at any given time. **Participants are encouraged to harvest the tweets as soon as possible upon receipt of the evaluation data after the decryption keys are released.** As such, to distinguish between no output due to deleted tweets from no output due to a system's inability to produce the results, each team is required to submit a file list along with their system output to indicate the source data availability. Even though this issue is only affected Twitter data, we ask teams to submit a list indicating the availability of all files in **set E** for ease of use. For consistency, use the file list distributed with set E (called '`filelist.txt`') and add a new field to indicate the file availability.

```
<DocID><tab><Available>
```

For example:

```
DF_AOA_TUR_0000116_20140900 TRUE
SN_TWT_TUR_2221137_20141021-02   FALSE
```

7.5 Submission Requirements

Each team is required to participate in the constrained training condition and is encouraged to participate in the unconstrained training condition. One of the goals of the LoReHLT evaluation is to track system performance over time. As such teams are required to submit at least one ensemble per the training condition participated. An *ensemble* is defined to be a set of three submissions, one at each checkpoint. Each team must designate one primary ensemble for cross-team comparisons but may submit addition contrastive ensembles for intra-team comparisons. At each checkpoint, teams are required to provide a short description of their submissions after they upload their system output. At the conclusion of the evaluation, each team is required to submit a single more formal system description that covers their primary and contrastive ensembles for all tasks the team participated in. The final results will be released to teams who submit a system description. The system descriptions will be compiled into the workshop proceedings. Teams can download the template for the system description on the NIST LoReHLT16 website.

Refer to the task specific sections below for the requirements on how to package the system output for a given task into a submission file.

8 Evaluation Rules and Requirements

The evaluation is an open evaluation where the test data is sent to the participants who will process and submit the output to NIST. As such, the participants have agreed to process the data in accordance with the following rules:

- The participant agrees not to investigate the evaluation data. Both human/manual and automatic probing of the evaluation data is prohibited to ensure that all participating systems have the same amount of information on the evaluation data.
- The participant agrees to abide by the terms guiding the use of the native informant[9].
- The participant agrees to process at least the constrained training track for each of the selected tasks.
- The participant agrees to complete all three checkpoints to be considered a complete submission for each selected task and training track combination.
- The participant agrees to participate in the dry run exercise to ensure evaluation readiness.
- The participant agrees to attend a post-evaluation workshop to present and discuss his/her systems. Failure to attend the workshop may result in participant being denied from participating in future evaluations.
- The participant agrees to the rules governing the publication of the results.

[9] contact NIST at lorehlt_poc@nist.gov if this presents a problem

9 Guidelines for Publication of Results

This evaluation follows an open model to promote interchange with the outside world. At the conclusion of the evaluation cycle, NIST will create a report that documents the evaluation. The report will be posted on the NIST web space and will identify the participants and the scores from various metrics achieved for task.

The report that NIST creates should not be construed or represented as endorsements for any participant's system or commercial product, or as official findings on the part of NIST or the U.S. Government.

9.1 Rules Governing Publication of Evaluation Results

The rules governing the publication of the LoReHLT evaluation results are similar to those used in other MIG evaluations.

- Participants are free to publish results for their own system, but participants must not publicly compare their results with other participants (ranking, score differences, etc.) without explicit written consent from the other participants.
- While participants may report their own results, participants may not make advertising claims about winning the evaluation or claim NIST endorsement of their system(s). Per U.S. Code of Federal Regulations (15 C.F.R. § 200.113): *NIST does not approve, recommend, or endorse any proprietary product or proprietary material. No reference shall be made to NIST, or to reports or results furnished by NIST in any advertising or sales promotion which would indicate or imply that NIST approves, recommends, or endorses any proprietary product or proprietary material, or which has as its purpose an intent to cause directly or indirectly the advertised product to be used or purchased because of NIST test reports or results.*
- All publications must contain the following NIST disclaimer:

  NIST serves to coordinate the evaluations in order to support research and to help advance the state- of-the-art. NIST evaluations are not viewed as a competition, as such results reported by NIST are not to be construed, or represented, as endorsements of any participant's system, or as official findings on the part of NIST or the U.S. Government.

10 Dry Run

All participants are required to participate in a dry run evaluation to demonstrate evaluation readiness. The purpose of the dry run is to exercise the evaluation infrastructure, not testing systems' ability to handle a new language. As such, the dry run intends to be flexible and at the same time to follow the protocol of the official evaluation as much as possible. Some of the differences between the dry run and the official evaluation are:

- Shorter time duration between checkpoints

- Time with native informant is for one hour[10] per team per task between IL Announcement and Checkpoint 1. It is the teams' responsibility to contact the data provider (Appen) to arrange time with the native informant during this period.
- The identity of the language is known before the IL Announcement (Mandarin).

11 LoReHLT Schedule

Milestone	***Date***
Initial version of evaluation plan published	Dec 11, 2015
Registration period	Feb 19 – May 16
6-month PI meeting in San Antonio, TX (LORELEI performers only)	Feb 24 – 26
Dry run period (see dry run schedule in Section 11.1)	Jun 01 – 10
Official evaluation period (see official evaluation schedule in Section 11.2)	Jul 05 – Aug 03
1.5-day NIST post-evaluation workshop co-located with DARPA PI meeting	Aug 28 – 29
2.5-day DARPA PI meeting (LORELEI performers only)	Aug 29 – 31

11.1 Dry Run Schedule

Milestones	***Date***
Encrypted data released by LDC	Jun 01
IL Announcement - Decryption keys for **set 0** and **set E** distributed by NIST - System description submission opens - Access to Native Informant begins - Submission for checkpoint 1 opens	Noon EDT Jun 02
Evaluation Checkpoint 1 - Access to Native Informant ends - Submission for checkpoint 1 closes - Decryption key for **set 1** distributed after submission made - Submission for checkpoint 2 opens	Noon EDT Jun 6
Evaluation Checkpoint 2 - Submission for checkpoint 2 closes - Decryption key for **set 2** and **set S** distributed after submission made - Submission for checkpoint 3 opens	Noon EDT Jun 08
Evaluation Checkpoint 3 - Submission for checkpoint 3 closes	Noon EDT Jun 10
System description submission closes	Noon EDT Jun 13
System description reviewed by NIST	Jun 14
Preliminary results released if system description is received	Jun 15
Native Informant Timeline (time amount is per team per task)	
1 hour between noon EDT Jun 02 to noon EDT Jun 06	

[10] Pending contract award prior to the dry run

11.2 Official Evaluation Schedule

Milestones	***Date***
Encrypted data released by LDC	Jul 05
IL Announcement - Decryption keys for **set 0** and **set E** distributed by NIST - System description submission opens - Access to Native Informant begins - Submission for checkpoint 1 opens	Noon EDT Jul 06
Evaluation Checkpoint 1 - Submission for checkpoint 1 closes - Decryption key for **set 1** distributed after submission made - Submission for checkpoint 2 opens	Noon EDT Jul 13
Evaluation Checkpoint 2 **-** Access to Native Informant ends - Submission for checkpoint 2 closes - Decryption key for **set 2** and **set S** distributed after submission made - Submission for checkpoint 3 opens	Noon EDT Jul 20
Evaluation Checkpoint 3 - Submission for checkpoint 3 closes	Noon EDT Aug 03
System description submission closes	Noon EDT Aug 09
System description reviewed by NIST	Aug 10
Preliminary results released if system description is received	Aug 11
Native Informant Timeline (time amount is per team per task)	
Up to 1 hour between noon EDT Jul 06 to noon EDT Jul 13 Up to 5 hours between noon EDT Jul 13 to noon EDT Jul 20 (or 4 hours if 1 hour was used between Jul 06 and Jul 13)	

12 Machine Translation (MT) Evaluation Specifications

12.1 Task Definition

Given a text document in the incident language, the MT system is required to automatically translate the document's content into English. The entire test set must be translated, even though only a subset of it will be scored in the machine translation evaluation.

12.2 Performance Measurements

BLEU and METEOR will be the primary metrics in Phase 1. BLEU and METEOR scores will be calculated at each checkpoint. Scoring will be done against four human reference translations. Scoring will be done preserving case. Other normalizations may be implemented for scoring purposes as necessary for the domains and data encountered.

NIST will investigate additional automatic approaches geared towards measurement of successful translation of content.

12.3 System Output Format

MT systems are required to output the translation conforming to the lorehlt-mt-v1.2.dtd[11]. A sample MT system translation file is given below:

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE mteval SYSTEM "lorehlt-mt-v1.2.dtd">
<mteval>
  <tstset>
    <doc docid="NW_ARX_UZB_164780_20140900">
      <seg id="segment-0">Who did vaccinations first?</seg>
      <seg id="segment-1">Go to navgation, search</seg>
      …
    </doc>
  </tstset>
</mteval>
```

The value of each `doc docid` attribute or `seg id` attribute must match exactly that used in the original LTF file.

Note that there is one MT system output file for each MT system input file, and the output file must have the same name as the input file.

12.4 System Submission Format

The MT system output files as described in 12.3 along with the file list as described in section 7.4 named '`filelist.txt`' should be placed into flat-file hierarchy and compressed into a .tgz or .zip file. There are no restrictions on the submission file name besides the suffix '`.tgz`' or '`.zip`'.

13 Situation Frame (SF) Evaluation Specifications

13.1 Task Definition

Given a text document in the incident language, an SF system is required to automatically identify the 0 or more situation frames covered in the document. Each system-generated SF consists of a situation type, place localization, and (for some types) status variables.

- Situation Type: A situation frame must be labeled as one of the pre-defined types in the LDC's "Annotation Guidelines for LORELEI Situation Frames"[12]. There are two general classes of situations: situations involving a 'need' (e.g., food supply, evacuation, etc.) or situations involving an 'issue' (e.g., civil unrest, terrorism, etc.). Regardless of the general class, the SF system will return a string for the situation type and a confidence score.
    - **SFType**: a text string indicating the enumerated type of situation.
    - **TypeConfidence**: a numeric confidence value indicating the strength of evidence supporting the identified situation type for the SF. (NOTE: TypeConfidence will not be evaluated during the 2016 evaluation.)

[11] ftp://jaguar.ncsl.nist.gov/lorehlt16/lorehlt-mt-v1.2.dtd

[12] "Annotation Guidelines for LORELEI Situation Frames"

- Place Mention: A situation occurs at a physical place, either a location or region. The SF system will identify the a named entity mention, in terms of the character extent and entity type, where the situation takes place if the document contains a named entity mention. In the event there is no named mention in the document, the system is expected to not return a mention. Reference SFs will be scored regardless of the 'Proxy' tag for place annotation.
    - **<u>Begin</u>**: Starting character offset of the mention within the source document
    - **<u>End</u>**: Ending character offset of the mention within the source document
    - **<u>EntityType</u>**: The entity type for the mention, either GPE or LOC. (NOTE: EntityType will not be evaluated during the 2016 evaluation.)
- Status Variables: Status variables indicate relevant context describing the situation.
    - The 'issue' situation types are not accompanied by status variables.
    - The 'need' situation types are accompanied by three status variables for each SF: "Need", "Relief", and "Urgency". The fill of each status variable is limited to an enumerated set prescribed by the annotation document. The system SF will list the following fills
        - **<u>Need</u>**: One of "Current", "Future only", "Past only"
        - **<u>Relief</u>**: One of "Sufficient", "Insufficient/Unknown sufficiency", "No known resolution"
        - **<u>Urgency</u>**: true | false

The entire test set must be processed even though only a subset of documents will be scored in the SF evaluation. Systems must provide the SFType to be evaluated. Systems specifically not addressing the geographic localization and/or status variables will not be evaluated with respect to the omitted fields.

13.2 Performance Measurements

The conceptual use of SF technology is to support down-stream applications that aggregate SF outputs to provide situational awareness using a variety of data sources that differ substantially with respect to the density of SFs and that simultaneously provides detailed supporting information about the situation. Thus, systems must directly support both low and high false alarm application scenarios and high quality supporting information.

This initial SF evaluation will not address the aggregation test case directly. Rather, system performance will be measured by their ability to correctly identify the right number of SFs using SF equivalency classes to assess performance at several levels of granularity while using a single system output. The assessment procedure will also not require systems to perform within-document entity co-reference by not penalizing a system for generating multiple SFs that identify mentions of the same reference entity.

In order to evaluate system performance, the following procedure will be performed for each document, for each entity type:

- Define the **equivalency class(es)** for the given metric:

    - The classes will describe which SF components to collapse in order to reduce the set of system frames. For example:
    - /place=place, need=*, relief=*, urgency=*/ treats SFs with differing status variables as equivalent.
    - /place=*, need=*, relief=*, urgency=*/ treats SFs with differing place and status variables as equivalent.
- Build the reduced set of scorable reference SFs (*R'*) using the equivalency classes and removing SFs with a 'true' proxy tag.
- Build the reduced set of scorable system SFs (*S'*) using the equivalency classes.
- Tally:
    - *Cor* = Correct SFs, the set of elements in R' with at least one matching S' based on the equivalency classes. Note: the definition of 'correct' is described below for each measure.
    - *Spu* = Spurious SFs, the set of elements in S' not matching any R' elements
    - *Del* = Deleted SFs, the set of elements in R' with no matching S' elements

The following metrics will be computed for the SFType, Place Mention, and Status Variables.

13.2.1 SFType Performance Measure

SFType performance will be measured as a 'recognition' task using Situation Frame Error (SFE) rate. The measure will answer: "Did the system produce the right 'type' of SFs for the document"? SFE is the ratio of spurious and deleted SFs to the number of reference SFs pooled over the test collection. For SFType performance, place mention and status variables for both system and reference SFs will be treated as equivalent.

Equivalence classes: /place=*, need=*, relief=*, urgency=*/

Correct SF requirements: The SFType of both system and reference SFs must match.

$SFE_{SFType} = |Spu + Del|/|R'|$

SFE will be calculated and reported over the full test collection, genre, SFType(s), Need SFType(s), and Issue SFTypes(s).

13.2.2 SFType+Place Mention Performance Measure

Joint SFType and Place Mention performance will be measured as Situation Frame Error (SFE) rate. The measure will answer: "Did the system produce the right set of 'type+place' SFs for the document"?. A system will not be penalized by creating multiple SFs for the same reference entity so long as the types match and the system's place mention extent matches at least one mention extent of the reference entity's mentions effectively 'no-scoring' the duplicates. For this measure, all status variables are treated as equivalent.

Equivalence classes: /place=place, need=*, relief=*, urgency=*/

Correct SF requirements: The SFType of both system and reference SFs must match and the system mention extent must match at least one mention of the reference entity's mentions.

$SFE_{SFType+Place} = |Spu + Del|/|R'|$

13.2.3 SFType+Place+Status Performance Measure

Joint SFType, Place Mention, and Status performance will be measured as Situation Frame Error (SFE) rate. The measure will answer: "Did the system produce the right set of 'type+place+status variable X' SFs for the document"?. Each status variable will be evaluated separately (even though 'need' and 'urgency' are inter-related) using a separate equivalence class for each variable and applying the same place mention matching rules as above.

Type+Place+Need:

Equivalence classes: /place=place, need=need, relief=*, urgency=*/

Correct SF requirements: The SFType, place mention (as described in 13.2.2), and Need status of both system and reference SFs must match

$SFE_{SFType+Place+Need} = |Spu + Del|/|R'|$

Type+Place+Relief:

Equivalence classes: /place=place, need=*, relief=relief, urgency=*/

Correct SF requirements: The SFType, place mention (as described in 13.2.2), and Relief status of both system and reference SFs must match

$SFE_{SFType+Place+Relief} = |Spu + Del|/|R'|$

Type+Place+Urgency:

Equivalence classes: /place=place, need=*, relief=*, urgency=urgency/

Correct SF requirements: The SFType, place mention (as in 13.2.2), and Urgency status of both system and reference SFs must match

$SFE_{SFType+Place+Urgency} = |Spu + Del|/|R'|$

13.3 System Output Format

The system output structure is a JSON structure and should confirm to the json schema "lorehlt-sf_output-schema_v0.2.json" that is available online[14]. Contained below is an initial example that is also available online[15].

[14] ftp://jaguar.ncsl.nist.gov/lorehlt16/lorehlt-sf_output-schema_v0.2.json

[15] ftp://jaguar.ncsl.nist.gov/lorehlt16/lorehlt-sf_sample-system-output_v0.2.json

```
[
  { "DocumentID": "123",
    "Type": "Water Supply",
    "TypeConfidence": 0.5,
    "PlaceMention": {
      "EntityType": "GPE",
      "Start": 25,
      "End": 40
    },
    "Status": {
      "Need": "Current",
      "Relief": "No known resolution",
      "Urgent": true
    }
  },
  { "DocumentID": "123",
    "Type": "Civil Unrest or Wide-spread Crime",
    "TypeConfidence": 0.7,
    "PlaceMention": {
      "EntityType": "LOC",
      "Start": 12,
      "End": 23
    }
  }
]
```

13.4 System Submission Format

The SF system output files as described in 13.3 named ‘system_output.json’ along with the file list as described in section 7.4 named ‘filelist.txt’ should be placed into flat-file hierarchy and compressed into a .tgz or .zip file. There are no restrictions on the submission file name besides the suffix ‘`.tgz`’ or ‘`.zip`’.

14 Named Entity Recognition (NER) Evaluation Specifications

14.1 Task Definition

Given a document in the incident language, an NER system is required to automatically identify and classify entity mentions into pre-defined entity types. Note only named mentions are targeted. The entity types in LORELEI/LORE tasks are listed as follows: (To be aligned with LDC NER annotations, we are planning to follow their definitions, so the following definitions may subject to change. A pointer to LDC’s annotation guidelines will be given later.)

- Person (PER): Person entities are limited to humans identified by name, nickname or alias.
- Geo-political Entity (GPE): GPE entities are composite entities, such as cities, provinces, countries, meaning there are several criteria that must be present to make something a GPE. GPEs consist of (1) a physical location, (2) a government, and (3) a population. All three of these elements must be present for an entity to be tagged as a GPE, as in: United States, China, Pennsylvania, Philadelphia
- Organization (ORG): Organization entities are limited to corporations, institutions, government agencies and other groups of people defined by an established organizational structure.

- Location (LOC): Location entities include geographically defined places such as bodies of water, mountains, etc.). Locations also include man-made structures like airports, highways, streets, factories and monuments (namely, facilities defined in previous programs such as ACE and KBP EDL).

Other types of named entities like events, animals, inanimate objects and monetary units will not be annotated.

14.2 Performance Measurements

Scoring metrics from TAC KBP2014/2015 tasks will be extended to the NER tasks. System output will be computed against the gold annotation output for precision (P), recall (R) and their balanced harmonic mean (F1). The official metric will be based on exact mention boundary matches, that is, a name mention is correctly labeled if its entity type and start/end offsets match those of a reference name mention. Specifically, we report these three metrics (P, R and F1) for strong_typed_mention from TAC2015 EDL measurements. The detailed description of TAC EDL scoring metrics is in section 2.2 in the overview paper: http://nlp.cs.rpi.edu/paper/kbp2015.pdf.

In addition to the exact match metric, we award systems for partial matches according to the degree of character overlap between system and key names for diagnostic analysis. The partial match scoring algorithm has two parameters: the recall overlap strategy and the precision overlap strategy.

- The per-name recall score of a name in the answer key is the fraction of its characters which overlap with the system name set according to the recall overlap strategy parameter. For the "MAX" strategy, this will be the characters overlapping with the single system name with maximum overlap. For the "SUM" strategy, this will be the number of its characters which overlap with any system mention.
- The recall score for a system is the mean of the per-name recall scores for all names in the answer key.
- The per-name precision score of a name in the answer key is the fraction of its characters overlapped by the reference set, where ”overlapping" is determined by the precision overlap strategy in the same manner as above for recall.
- The precision score for a system is the mean of the per-name precision scores for all names in the answer key.

We will report scores for all four parameter combinations. The scorer is available at: https://github.com/wikilinks/neleval.

14.3 System Output Format

An NER system is required to automatically generate an output file, which contains one line for each mention, where each line has the following tab-delimited fields. Please note that while the format is identical to that of TAC2014/2015 EDL, some fields will just be placeholders as noted below. Using the same format eliminates needs for making changes to the scorer code. Besides, full EDL is expected in year 2 and beyond.

```
Field1<tab>Field2<tab>Field3<tab>...<tab>Field8
```

where:

Field 1: system run ID, unique team_id to identify each team and their runs

Field 2: mention ID, unique for each entity name mention

Field 3: mention head string, the full head string of the entity mention

Field 4: document ID: mention head start offset – mention head end offset, an ID for a document in the source corpus from which the mention head was extracted, the starting offset of the mention head, and the ending offset of the mention head.

Field 5: NIL (in future this is a place holder for reference KB link entity ID)

Field 6: entity type: {GPE, ORG, PER, LOC} type indicator for the entity

Field 7: all should be of type {NAM}

Field 8: a confidence value. Set to 1 for Phase I. In the future this field will be replaced by entity linking confidence score.

Sample NER output:

```
NIST  NW_ARX_UZB_164780_20140900-NE1  Eduard Jennerni  NW_ARX_UZB_164780_20140900:479-493   NIL  PER  NAM  1.0
NIST  NW_ARX_UZB_164780_20140900-NE2  Glostershir      NW_ARX_UZB_164780_20140900:614-624   NIL  LOC  NAM  1.0
NIST  NW_ARX_UZB_164780_20140900-NE3  Eduard Jenner    NW_ARX_UZB_164780_20140900:1038-1050 NIL  PER  NAM  1.0
NIST  NW_ARX_UZB_164780_20140900-NE4  Jenner           NW_ARX_UZB_164780_20140900:1365-1370 NIL  PER  NAM  1.0
...
```

Note that there is only one NER system output file, and the output file can have any name as long as it has a “.tab” extension.

14.4 System Submission Format

The NER system output file as described in 14.3 along with the file list as described in section 7.4 named ‘filelist.txt’ should be placed into flat-file hierarchy and compressed into a .tgz or .zip file. There are no restrictions on the submission file name besides the suffix ‘.tgz’ or ‘.zip’.